Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Reinforcement-Learning-based Optimization for Day-ahead Flexibility Extraction in Battery Pools ⋆

Georgios C. Chasparis * **Christian Lettner** *

* *Software Competence Center Hagenberg GmbH, Softwarepark 21, 4232 Hagenberg, Austria (e-mail: {georgios.chasparis,christian.lettner}@scch.at)*

Abstract: We address the problem of trading energy flexibility, derived from pools of residential Photovoltaic and battery-storage systems, to the Day-ahead electricity market. By flexibility, we imply any additional energy that can be stored to or withdrawn from the participating batteries/households at a given time during the next day. The optimization variables include the selection/activation of a subset of participating batteries and the amount of flexibility that should be extracted. Furthermore, the optimization objective corresponds to the expected forecast revenues that can be generated by trading this flexibility to the Day-ahead electricity market. Given the high computationally complexity of a full scale optimization in the case of a large number of participating batteries, we propose a reinforcement-learning-based methodology, which admits linear complexity with the number of participating batteries. The proposed methodology advances prior work with respect to the integration of a large number of batteries. Furthermore, it extends prior work of the authors with respect to providing analytical performance guarantees in comparison with the baseline/nominal operation of the battery. Finally, we compare through simulations the performance of the proposed method with a Linear-Programming-based optimization that provides the exact optimum.

Keywords: Control of renewable energy resources; Smart grids; Approximate dynamic programming; Day-ahead spot electricity market

1. INTRODUCTION

Recently, the number of Photovoltaic (PV) and battery-storage systems in residential buildings constantly increases, Kairies et al. (2016). So far, such storage systems are mainly used to maximize the on-site absorption of the PV generation. Given the current need for increasing the percentage of renewable energy fed into the grid, the available charging/discharging flexibility can also be used to react to price variations in the Day-ahead (DA) or the Intra-day (ID) electricity spot-market. The representative agent (or aggregator) could make such decisions over the specific use of the storage units with respect to the optimal participation at the DA and ID spot markets, the benefits of which can then be transferred to the owners of the participating batteries.

In this paper, we focus on addressing the problem of optimal activation of a set of residential battery-storage systems in the DA spot market. The underlying assumption is that an aggregator directly controls a set of batteries, thus any charging/discharging flexibility potential can be extracted in real-time. The proposed scheme will be based upon an approximate-dynamic-programming (or reinforcement-learning) methodology. According to this scheme, an approximation function of the performance is being trained (using historical data) that can be used to generate optimal biddings/schedules for the DA market. By design, the proposed scheme is flexible enough to accommodate the possibility of erroneous forecasts as well as the need for re-optimizing in real-time upon receipt of corrected/updated forecasts. The main novelty of the proposed methodology lies on the possibility of incorporating a large number of batteries, while we also provide analytical performance guarantees.

The remainder of this paper is organized as follows. In Section 2, we present related work and the main contributions of this paper. Section 3 provides the main framework and challenges of this problem. Section 4 discusses the optimization problem and objective and Section 5 presents a reinforcement-learning-based scheme that it is specifically tailored for this class of problems. Section 6 presents an evaluation of the proposed framework on real-world data and a comparative analysis with a linear-programming-based methodology that provides the exact optimum. Finally, Section 7 provides concluding remarks.

2. RELATED WORK AND CONTRIBUTIONS

Demand response is either *commitment-based*, where consumers agree on reducing the load during peak hours Ruiz et al. (2009), Chen et al. (2014) or *incentive-based*, where financial incentives are offered to the consumers Herter (2007), Triki and Violi (2009), Xu et al. (2016). For example, a *commitment-based approach* has been proposed by

⋆ This work has been supported by the Austrian Research Agency FFG through the research project Flex+ (FFG # 864996).

Chen et al. (2014), where the operator distributes portions of its desired aggregated demand to the households, using an average consensus algorithm. On the other hand, an *incentive-based approach* has been proposed by Xu et al. (2016), where each participating household communicates to the operator a bidding curve, that is a function that provides the load adjustment that each user is willing to perform at a given price.

Apart from these approaches, there is an alternative methodology which can be considered as a combination of the two and it is closer to the one employed in this paper. According to such methodology, an aggregator directly extracts the required flexibility from the participating equipment when necessary. In return the aggregator offers to the owners of the equipment an agreed financial compensation. Such methodology is usually referred to as *demand-response aggregation* Parvania et al. (2013). It has been employed in Parvania et al. (2013), where aggregators can activate load reduction in a set of consumers according to an agreed demand-response strategy for each consumer. Similar in spirit is also the work in references Iria et al. (2017); Nan et al. (2018), where an aggregator directly controls a set of different types of loads in residential buildings to reduce total electricity consumption. As expected, a feature that distinguishes *demand-response aggregation* is the self-scheduling or activation optimization problem, that is the optimization of optimally utilizing the available flexibility (stemming from several households) over a future time horizon. Such feature (of multiple households) is not usually considered in the context of participation in a wholesale electricity market (see, e.g., Gomez-Villalva and Ramos (2003); Philpott and Pettersen (2006)).

In the context of battery-storage systems, *demand-response aggregation* (as discussed in the previous paragraph) has not yet been addressed in an effective and computationally efficient way. In this context, the aggregator wishes to compute an optimal (day-ahead) schedule for extracting flexibility (charge, discharge or do nothing) for each one of the participating batteries. So far, such optimization problem has mostly been addressed for single battery systems, e.g., Mohsenian-Rad (2016); He et al. (2016). Existing methodologies also include a detailed modeling of the battery as well as a detailed description of the cycle costs of the battery due to the frequent charging/discharging He et al. (2016). It may include computations of optimal charging/discharging bids for the day-ahead electricity market, as in Mohsenian-Rad (2016); He et al. (2016), or the intra-day/hour-ahead electricity market, as in Jiang and Powell (2015). In order to effectively address the uncertainty of the initial/final stage-of-charge of the battery, Jiang and Powell (2015) also employs an Approximate Dynamic Programming (ADP) formulation.

In this paper, participation to the DA wholesale electricity market is implemented by directly controling the battery-storage systems, as in Mohsenian-Rad (2016); He et al. (2016); Jiang and Powell (2015). Recent work of the authors Chasparis et al. (2019) has proposed an ADP framework that can efficiently be employed for a large number of battery-storage systems. It extended prior work by addressing multiple battery-storage systems, contrary to the single battery-storage system in Mohsenian-Rad (2016), He et al. (2016), and Jiang and Powell (2015). In comparison with Chasparis et al. (2019), this paper presents an improved design that provides analytical guarantees over the long-term performance in comparison to the baseline/nominal operation of the battery. Furthermore, in the case of a single battery, a comparison is performed through simulations with a linear-programming-based optimization that provides the exact optimum.

3. FRAMEWORK AND CHALLENGES

Time is divided into intervals ΔT, that define the instances at which measurements are collected and decisions are revised regarding the operation of the battery. Throughout the paper, we will assume that $\Delta T = 1/4h$, which implies that each day is divided into $N = 96$ time intervals. In several cases, we will interchangeably use the time variable t to also denote the index of the corresponding time interval. Thus, $t+1$ will often denote the next time interval.

We are provided with a set $\mathcal{I}$ of households that are equipped with PV panels and battery-storage systems. Let also i be a representative element of this set. At any given time interval t each of these households can be characterized by the electrical power generated from the PV panels $P_{\mathrm{PV},i}(t) \geq 0$, the electrical load consumed by the users/residents of the household $P_{\mathrm{load},i}(t) \geq 0$, and the state-of-charge $\mathrm{SOC}_i(t)$ of the battery. Given the small duration of these time intervals (15min), all power variables (such as, $P_{\mathrm{PV},i}(t)$ and $P_{\mathrm{load},i}$) will always be defined as the corresponding mean value over the current time interval t. Thus, $P_{\mathrm{PV},i}(t)$ and $P_{\mathrm{load},i}(t)$ will assume a constant value over the time interval t. On the other hand, any energy variable, e.g., $E_{\mathrm{PV}}(t) = P_{\mathrm{PV},i}(t)\Delta T$, will denote the total energy exchanged during the time interval t. Finally, the $\mathrm{SOC}_i(t)$ will correspond to the state-of-charge at the beginning of time interval t.

In several cases, we will also use the notation $\Delta P_i(t) = P_{\mathrm{PV},i}(t) - P_{\mathrm{load},i}(t)$ to denote the *excess power* during time-interval t (which can be positive or negative). Also, let $P_{\mathrm{g},i}(t)$ denote the power received from the grid at time interval t, and $P_{\mathrm{b},i}(t)$ the power that flows to the battery (before any charging/discharging losses apply). At any given interval t, the power balance in the household dictates:[1] $P_{\mathrm{g},i}(t) = P_{\mathrm{b},i}(t) - \Delta P_i(t)$.

3.1 Baseline battery operation

Each of these batteries is equipped with its own (logical) controller that considers *autarky* as the main priority. According to such baseline controller, a) if $\Delta P_i(t) > 0$, then the battery is charged and, if full, the excess power is fed into the grid, and b) if $\Delta P_i(t) < 0$, then the battery is discharged and, if empty, additional power is procured from the grid.

At the beginning of each time interval t, and given $\mathrm{SOC}_i(t)$ of battery i, as well as the current excess power $\Delta P_i(t)$, we can compute the *baseline* power to the battery (i.e., the power to the battery under the standard, *autarky-based controller*). This is a straightforward calculation that is

[1] For the sake of clarity of presentation, we neglect the (generally small) energy losses in the AC-DC inverter.

based upon several features of the battery (such as the maximum charging/discharging power, the capacity and the charging/discharging loss rates of the battery). The details of such computation can be found at (Chasparis et al., 2019, Algorithm 1). We will denote this computation by the following function:

$$P_{\mathrm{b,base},i}(t) = \mathcal{B}_i(\Delta P_i(t), \mathrm{SOC}_i(t)).$$

Similarly, we may define the corresponding baseline power from the grid, denoted by $P_{\mathrm{g,base},i}(t)$.

3.2 Energy flexibility potential

The *charging* and *discharging energy (flexibility) potential* at time interval t of household i refers to the amount of energy that the household may additionally procure from and feed into the grid during time interval t, respectively. In order to accurately compute the energy potentials, we need to take into account the current baseline operation of the battery at time interval t. The quantities $v_{c,i}(t) \geq 0\,, v_{d,i}(t) \leq 0$ will denote the *charging and discharging potential that is available in household* i, respectively. The charging potential is defined as: [2]

$$v_{c,i}(t) \doteq \left[P^*_{\mathrm{b},c,i}(t) - P_{\mathrm{b,base},i}(t)\right]_+ \Delta T, \qquad (1)$$

where $P^*_{\mathrm{b},c,i}(t) \geq 0$ is the maximum possible (mean) charging power to the battery. Similarly, the discharging potential is defined as:

$$v_{d,i}(t) \doteq \left[P^*_{\mathrm{b},d,i}(t) - P_{\mathrm{b,base},i}(t)\right]_- \Delta T \qquad (2)$$

where $P^*_{\mathrm{b},d,i}(t) \leq 0$ is the maximum (in absolute value) mean discharging power from the battery. We will briefly express the above computations of the energy potential by

$$[v_{d,i}(t), v_{c,i}(t)] = \mathcal{V}_i(P_{\mathrm{b,base},i}(t), \mathrm{SOC}_i(t)).$$

In order to better understand the notions of charging/discharging flexibility potential, let us consider the $\mathrm{SOC}_i(t)$ schematic profile of a battery in Figure 1. Specifically, let us consider the first time interval ΔT on the left-hand side. Given that $\Delta P_i(t) > 0$ during this time interval, i.e., there is a positive excess of energy, we should expect that under the baseline operation of the battery, the battery will be charged and $P_{\mathrm{b,base},i}(t) > 0$. The energy that would be charged to the battery due to the baseline operation has been highlighted with the yellow shaded area. This implies that the available charging potential on that interval will be the remaining green shaded area. On the other hand, the positive excess energy $\Delta P_i(t)\Delta T$, together with the energy available in the battery (red shaded area), corresponds to the discharging potential.

3.3 Control variables and system dynamics

In this work, we will be concerned with activating or committing part of the available charging/discharging potential to the DA electricity market. The parameter $u_i(t) \in [-1, 1]$ will denote the *activation factor* of battery i at time interval t. If $u_i(t) \geq 0$, then $|u_i(t)|$ represents the *fraction of the charging potential that is activated.* Analogously, if $u_i(t) \leq 0$, then $|u_i(t)|$ represents the *fraction of the discharging potential that is activated.*

[2] We use the notation $[x]_+ \doteq \max\{x, 0\}$ and $[x]_- \doteq \min\{x, 0\}$.

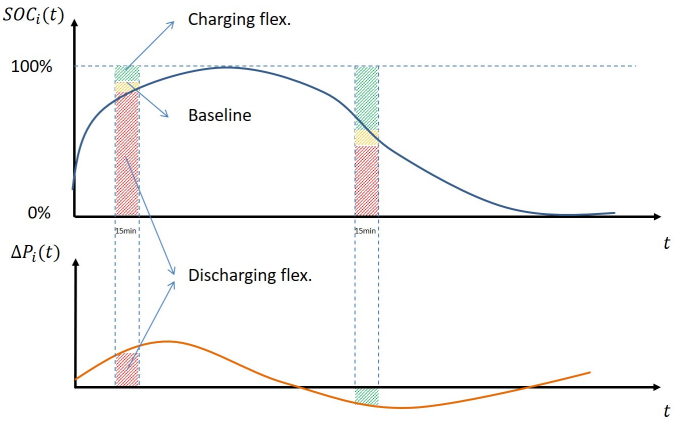

Fig. 1. Example of charging/discharging flexibility potential over a time interval ΔT.

We will define $E_i(t)$ as the energy traded/committed to the DA market during time interval t. We will also adopt the convention that the energy is positive if it is charged to the household/battery and negative otherwise. In other words, if $u_i(t) \geq 0$, then $E_i(t) = u_i(t)v_{c,i}(t) \geq 0$ (*energy is charged to household* i), and if $u_i(t) \leq 0$, then $E_i(t) = -u_i(t)v_{d,i}(t) \leq 0$ (*energy is discharged from household* i). In several cases, we will also denote $E(t)$ to be the total energy charged to/discharged from $\mathcal{I}$.

Optimal activations over the duration of a future horizon require an explicit knowledge of how the flexibility potential varies due to prior activations. In particular, and given the previous definitions, the flexibility potential can recursively be computed for each time interval t by executing the following recursions in sequence.

$$\begin{aligned}
\mathrm{SOC}_i(t+1) &= \Sigma_i(\mathrm{SOC}_i(t), P_{\mathrm{b,base},i}(t), u_i(t)) \\
P_{\mathrm{b,base},i}(t+1) &= \mathcal{B}_i(\Delta P_i(t+1), \mathrm{SOC}_i(t+1)) \\
[v_{d,i}(t+1), v_{c,i}(t+1)] &= \mathcal{V}_i(P_{\mathrm{b,base},i}(t+1), \mathrm{SOC}_i(t+1))
\end{aligned}$$

The mapping Σ_i provides the $\mathrm{SOC}_i(t+1)$ at the beginning of the next time interval, given the previous state-of-charge, the baseline power to the battery, and the activation. We may view the state-of-charge $\mathrm{SOC}_i(t)$ and the baseline power to the battery $P_{\mathrm{b,base},i}(t)$ as internal states of the above system dynamics, and the flexibility potentials $v_{d,i}(t), v_{c,i}(t)$ as the outputs of this system. The state equations are sub-linear, as presented in Chasparis et al. (2019).

3.4 DA-optimization and challenges

We assume that there is a *representative agent* (or *aggregator*) of the set of households/batteries $\mathcal{I}$ which has direct access to the operation of all batteries. We will briefly denote it by RA. For example, in the context of the Flex+ project [3], such representative agent or aggregator corresponds to a software platform that automatically computes the available flexibility and optimal activations.

The RA tries to exploit any market opportunities that may arise in the DA spot electricity market, due to variations in the electricity price. In this context, and at the beginning of each day, the RA has available (in the form of estimates) the initial state-of-charge of the next

[3] https://www.flexplus.at

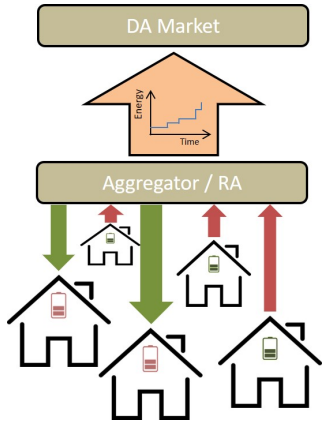

Fig. 2. Schematic of DA optimization framework. The RA trades energy directly in the DA spot electricity market through activation schedules of charging/discharging flexibility potential.

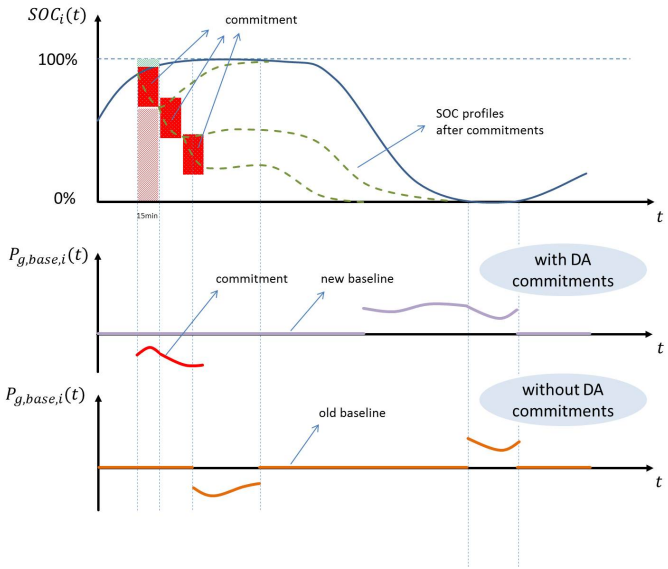

Fig. 3. Long-term negative impact of DA commitment schedules. An example is depicted of the differences in the baseline operation of the battery after feeding a discharging commitment into the grid (red line).

calendar day $\mathrm{SOC}_i(0)$, the time series of excess energy $\{\Delta P_i(t)\}_{t=0}^{96}$ for each battery $i \in \mathcal{I}$, and the time series of the DA electricity price $\{\lambda_{\mathrm{DA}}(t)\}_{t=0}^{96}$ over the duration of the next day.

Given that the DA electricity price forecast is only available for the following day or for the next two days, naturally an optimization horizon will be restricted to one or two days ahead. However, under such restricted optimization horizon, the exact impact of the day-ahead schedules on the long-term utility (over several days) might be unknown.

Figure 3 demonstrates such a scenario of negative long-term impacts due to myopically derived schedules. In this scenario, a one-day-ahead optimization may dictate some energy discharges, which are temporarily profitable, since they increase the household's revenues by feeding in energy. However, this decision significantly drops the state-of-charge of the battery. Thus, if in the following day(s), the load increases and the state-of-charge cannot cover for it, the household will need to procure additional energy from the grid. In other words, under the *new baseline* operation of the battery (which results from the activation of earlier energy commitments), the household increases its costs in the long term, as compared to the *original baseline* operation of the battery (without the activation of these commitments). Thus, short-term one-day-ahead optimization may have negative long-term impacts in the utility of the household (i.e., higher costs), which need to be taken into account. This complicates the optimization problem since we need to guarantee that decisions are also optimal in the long term.

4. OPTIMIZATION PROBLEM AND OBJECTIVE

Assuming a set of participating batteries $\mathcal{I}$, we are interested in maximizing the long-term utility received through the activation of day-ahead commitment schedules in the DA market. The difficulty emerges from the fact that although the optimization criterion expands over a large (or infinite) horizon, in reality decisions may only be made once per day. In fact, schedule bids may be submitted once per day for the DA market of the next day and cannot be revised during execution. Thus, *we are facing an infinite horizon optimization criterion that may only be addressed through repeated one day-ahead optimizations.*

The optimization variables are $\{u(1), u(2), ..., u(N)\} \in \mathcal{U} \doteq [-1,1]^{|\mathcal{I}|} \times ... \times [-1,1]^{|\mathcal{I}|}$, where $u(t) = [u_i(t)]_i$ is the vector of activations over the set of batteries $\mathcal{I}$. The discussion of Section 3.4 on the possible long-term impacts of the DA commitments suggests the following decomposition of the (instantaneous) utility/objective function

$$g(x,u) = g_{\mathrm{DA}}(x,u) + g_{\mathrm{BA}}(x,u). \tag{4}$$

The objective function depends on the current state variables $x \in \mathcal{X}$, as defined in Section 3.3, and the current activations u. The state variables x may also incorporate forecasts of future prices and excess available energy, which may help us make more informative decisions. This will be specified in a forthcoming section. We decompose the utility function into the $g_{\mathrm{DA}}(x,u)$ utility that includes the utility of the DA commitment schedules, and the $g_{\mathrm{BA}}(x,u)$ utility that captures the difference with the utility under the baseline operation. Note that the definition of $g_{\mathrm{BA}}(x,u)$ also requires a reference initial time and state based on which the baseline profile is defined.

We are facing an infinite horizon optimization criterion, which may only be addressed through repeated one-day-ahead optimizations, computed at the beginning of each day (DA bids) and executed during the following day. The infinite-horizon optimization criterion can be expressed as:

$$J_{\infty}^{\mu}(x_0) = \sum_{t=0}^{\infty} \delta^t g(x(t), \mu(x(t))), \quad x(0) = x_0 \tag{5}$$

where the process is initiated at x_0, for some discount factor $\delta \in (0,1)$. Decisions on activations are also given by $u(t) = \mu(x(t))$, where $\mu : \mathcal{X} \to \mathcal{U}$ is our stationary policy for generating activations.

In practice, such optimization will be based upon forecasts (of, e.g., DA prices, PV generation and non-flexible load),

which implies that the above expression should be written in terms of expectations. Although the forthcoming methodology can accommodate the possibility of erroneous forecasts, ***the presentation will focus only on the case of perfect forecasts***.

5. ZERO-STEP-CONTROL APPROXIMATE DYNAMIC PROGRAMMING

In this paper, we propose a methodology that is based on ADP. Such approach can capture long-term impacts on the utility through a large number of simulated scenarios. The complexity of the problem though as well as the rather limited amount of real-world data do not allow for considering a full-scale black-box approach. Instead, we will introduce a methodology that is tailored to the specifics of the problem, while providing performance guarantees over the long-term impact of the DA schedules, captured by the original objective (5).

For the sake of clarity of presentation, ***the forthcoming analysis will be restricted to a single battery***. An extension to multiple batteries will be straightforward. In this section and by abusing notation, the time index t may appear as a subscript of the state and control variable, e.g., x_t and u_t.

5.1 Notation and state variables

Let us first introduce the following notation:

- We define the *infinite-step utility function under the baseline operation* as
$$J_{\text{base}}(x_0) \doteq \sum_{t=0}^{\infty} \delta^t g(x_t, 0),$$
i.e., it corresponds to the infinite-step utility function when no commitments are offered to the DA spot market.
- We define the *k-step-control utility* as
$$J_{\infty}^{\mu,k}(x_0) \doteq \sum_{t=0}^{k} \delta^t g(x_t, \mu(x_t)) + \delta^k \sum_{t=k+1}^{\infty} \delta^{t-k} g(x_t, 0),$$
i.e., it corresponds to the infinite-step utility when for $t \leq k$ we employ policy μ, while for any time $t > k$ we follow the baseline controller.
- We define the *k-plus-L-step-control utility* as follows:
$$R_L^{\mu,k}(x_0) \doteq \sum_{t=0}^{k} \delta^t g(x_t, \mu(x_t)) + \delta^k \sum_{t=k+1}^{k+L} \delta^{t-k} g(x_t, 0),$$
where the baseline controller is applied for $L \geq 1$ steps after k. Under the baseline controller, we also define:
$$R_L^{\text{base},k}(x_0) \doteq \sum_{t=0}^{k+L} \delta^t g(x_t, 0).$$
- We define $\Delta R_L(x, \mu(x)) \doteq R_L^{\mu,0}(x) - R_L^{\text{base},0}(x)$, which captures the difference between the 0-plus-L-step-control utility with the corresponding one under the baseline controller.

The state x_t at the beginning of time interval t will comprise the following parameters for all batteries $i \in \mathcal{I}$:

- $\text{SOC}_i(t)$, state-of-charge of battery i at time t,
- $v_{c,i}(t)$, charging potential of i at time interval t,
- $v_{d,i}(t)$, discharging potential of i at time interval t,
- $\{\lambda_{\text{DA}}(t), \lambda_{\text{DA}}(t+1), ..., \lambda_{\text{DA}}(t+L)\}$, the L-length sequence of future DA prices,
- $\lambda_{\text{DA},f}(t)$, the average future price over the time horizon of L time steps ahead, i.e., within time intervals $t+1$ until $t+L$,
- $\psi_i^b(t) \in \mathbb{R}_+$, total energy that is procured from the grid by battery i under the baseline controller over a future time horizon of L steps, i.e., within time intervals $t+1$ until $t+L$,
- $\phi_i^b(t) \in \mathbb{R}_-$, total energy that is fed into the grid by battery i under the baseline controller over a future time horizon of L steps, i.e., within time intervals $t+1$ until $t+L$.

Finally, denote $\eta_{c,i}$, and $\eta_{d,i}$ to be the charging and discharging efficiency rates of battery i, respectively, and χ_i to be the energy capacity of battery i.

5.2 Zero-step-control guarantees

In this section, and given the above definitions, we establish long-term guarantees over the original criterion $J_{\infty}^{\mu}(x)$, and under certain conditions. For some $\varepsilon > 0$, let us define the *zero-step-control optimal policy* $\mu_{\varepsilon}^* : \mathcal{X} \mapsto [-1, 1]$ such that, for any state x we have

$$\mu_{\varepsilon}^*(x) \doteq \begin{cases} \arg \underbrace{\max_{u \in \mathcal{U}(x)} \{\Delta R_L(x,u) - \varepsilon\}}_{\doteq \times}, & \text{if } \times > 0 \\ 0 & \text{if } \times \leq 0 \end{cases} \tag{6}$$

where [4]

$$\mathcal{U}(x) \doteq \left[\frac{\max\{v_{d,i}(t), \eta_{d,i}\eta_{c,i}\phi_i^b(t)\}}{v_{d,i}(t)}, \frac{\min\{v_{c,i}(t), \psi_i^b(t)\}}{v_{c,i}(t)} \right].$$

This policy maximizes the 0-plus-L-step-control utility in comparison to the corresponding baseline utility starting from the same state x. Note that the optimal policy suggests a non-zero control $\mu_{\varepsilon}^*(x) \neq 0$ only if $\Delta R_L(x, \mu_{\varepsilon}^*(x)) > 0$, i.e., the 0-plus-$L$-control utility increases with respect to the corresponding baseline utility starting from x. We will impose the following design assumption on the definition of time horizon of L steps.

Assumption 5.1. The time horizon of L steps is sufficiently small such that for any time interval t: a) either $\{\phi_i^b(t) \leq 0, \psi_i^b(t) = 0\}$ or $\{\phi_i^b(t) = 0, \psi_i^b(t) > 0\}$ but not both, b) if $\psi_i^b(t) > 0$ then $\phi_i^b(t+1) = 0$, and c) if $\phi_i^b(t) < 0$, then $\psi_i^b(t+1) = 0$.

Essentially, Assumption 5.1 requires that the horizon of L steps ahead is short enough such that the household does not procure and feed-in energy concurrently within a window of L time steps.

Lemma 5.1. Let $L \geq 1$ satisfy Assumption 5.1. For some initial state x_0, let the control input sequence $\{u_0^*, 0, ..., 0\}$ be implemented, where at zero-step $u_0^* = \mu_{\varepsilon}^*(x_0)$ is applied, followed by an L-step implementation of the baseline controller. Let also $\{x_0, x_1^0, x_2^0, ..., x_L^0\}$ denote the evolution of the state under this control sequence, and

[4] In case $v_{c,i}(t) = 0$ or $v_{d,i}(t) = 0$ the corresponding fraction is assigned the 0 value.

$\{x_0, x_1^b, x_2^b, ..., x_L^b\}$ denote the corresponding sequence under the baseline operation. Then, $x_{L+1}^0 \equiv x_{L+1}^b$.

Proof. Without loss of generality, let us consider the case that $\{\phi_i^b(0) = 0, \psi_i^b(0) > 0\}$ within the upcoming L time steps. (The case of $\{\phi_i^b(0) \leq 0, \psi_i^b(0) = 0\}$ follows similar reasoning.) Recall that according to the baseline controller, the household procures energy from the grid at time t only if $\mathrm{SOC}_i(t) \equiv 0$ and $P_{\mathrm{load},i}(t) > 0$. Thus, $\psi_i^b(0) > 0$ implies that, under the baseline controller, there exists a time step indexed by $1 \leq t' \leq L$ within which the battery reaches the empty state. Let also $t' \leq t'' \leq L$ be the last time step within which the battery is at an empty state. Also,

$$0 \leq \psi_i^b(0) = -\mathrm{SOC}_i(0)\chi_i\eta_{d,i} - \Delta T \sum_{t=1}^{t''} \left\{ [\Delta P_i(t)]_- + [\Delta P_i(t)]_+ \eta_{c,i}\eta_{d,i} \right\} \quad (7)$$

i.e., the energy that household i procures from the grid during the upcoming t'' steps will be at least as much as the total energy needed during the same period (taking also into account the energy losses when charging/discharging the battery). Given our control decision $u_0^* \geq 0$, energy $0 \leq E_i(0) \doteq u_0^* v_{c,i}(0) \leq \psi_i^b(0)$ is procured within time interval 0. This implies that, under the control sequence $\{u_0^*, 0, ..., 0\}$, the battery will also reach the empty state within interval t'' or earlier. As a consequence, at time steps $t > t''$, where the baseline controller is implemented, the state will coincide with the corresponding state under the baseline controller profile. Thus, at the end of the Lth time interval, we should have $x_{L+1}^0 \equiv x_{L+1}^b$, which concludes the proof. □

Lemma 5.1 relates the state variables at the end of the zero-step-control sequence with the corresponding ones of the baseline process. This further allows for establishing performance guarantees when μ_ε^* is implemented repeatedly, as the following proposition demonstrates.

Proposition 5.1. Under the hypotheses of Lemma 5.1 and for a discount factor $\delta \in (0, 1)$, let the zero-step-control optimal policy μ_ε^*, $\varepsilon > 0$, be implemented repeatedly for all future time intervals $t = 0, 1, ...$, starting from any initial state $x_0 \in \mathcal{X}$. Then,

$$J_\infty^{\mu_\varepsilon^*}(x_0) \geq J_{\mathrm{base}}(x_0).$$

Proof. As in Lemma 5.1, let $x_0, x_1^b, x_2^b, ..., x_L^b, ...$ denote the evolution of the state under the baseline controller, and $x_0, x_1^0, x_2^0, ...x_L^0, ...$ denote the evolution of the state when implementing the optimal zero-step-control sequence $\{u_0^*, 0, 0, ...\}$, where $u_0^* = \mu_\varepsilon^*(x_0)$. Without loss of generality, consider the case that $\psi_i^b(0) > 0$ which will lead to a potential activation $0 \leq E_i(0) \doteq u_0^* v_{c,i}(0) \leq \psi_i(0)$. Given Lemma 5.1, we have $x_{L+1}^0 \equiv x_{L+1}^b$, $x_{L+2}^0 \equiv x_{L+2}^b$, ..., which implies that

$$J_\infty^{\mu_\varepsilon^*,0}(x_0) = g(x_0, u_0) + \delta g(x_1^0, 0) + ... + \delta^L g(x_L^0, 0) + \delta^{L+1} J_{\mathrm{base}}(x_{L+1}^b).$$

Given the definition of the optimal zero-step-control policy μ_ε^* in (6), we should also have:

$$J_\infty^{\mu_\varepsilon^*,0}(x_0) - J_{\mathrm{base}}(x_0) = R_L^{\mu_\varepsilon^*,0}(x_0) - R_L^{\mathrm{base},0}(x_0) = \Delta R_L(x_0, \mu_\varepsilon^*(x_0)) \geq 0.$$

Under the new profile of the states $x_0, x_1^0, x_2^0, ...x_L^0, ...$, let $\psi_i^0(1)$ denote the energy needed within the upcoming L time steps under the baseline controller when starting from state x_1^0. By Assumption 5.1, $\psi_i^0(1) \geq 0$ and $\phi_i^0(1) = 0$, i.e., the battery may not feed-in energy within interval $t = 2$ until $t = L + 1$. (If energy is fed-in before time interval $L + 1$, it contradicts the fact that under u_0^* the energy charged is less than the energy needed, i.e., the battery gets empty before $L + 1$.) It suffices to consider the case that $\psi_i^0(1) > 0$, which implies that we still need energy under the new baseline. Given that $u_0^* > 0$ (i.e., we charged energy) and $x_\tau^0 \equiv x_\tau^b$ for all $\tau \geq L + 1$, then $\psi_i^0(1) \leq \psi_i^b(1)$, i.e., we need less energy as compared to the original baseline. We conclude that under an updated control sequence $\{u_0^*, u_1^*, 0, ...\}$, where the state evolves as $\{x_0, x_1^0, x_2^1, x_3^1, ..., x_L^1, x_{L+1}^1\}$, we have $x_{L+2}^1 \equiv x_{L+2}^b$, i.e., the state merges again to the baseline profile. In this case,

$$\begin{aligned}
&J_\infty^{\mu_\varepsilon^*,1}(x_0) - J_{\mathrm{base}}(x_0) \\
&= R_L^{\mu_\varepsilon^*,1}(x_0) - R_L^{\mathrm{base},1}(x_0) \\
&= g(x_0, u_0^*) + \delta R_L^{\mu_\varepsilon^*,0}(x_1^0) - R_L^{\mathrm{base},1}(x_0) \\
&= g(x_0, u_0^*) + \delta R_L^{\mathrm{base},0}(x_1^0) + \delta \Delta R_L(x_1^0, u_1^*) - R_L^{\mathrm{base},1}(x_0) \\
&= R_L^{\mu_\varepsilon^*,0}(x_0) + \delta^{L+1} g(x_{L+1}^b, 0) + \delta \Delta R_L(x_1^0, u_1^*) - \\
&\quad R_L^{\mathrm{base},1}(x_0) \\
&= R_L^{\mu_\varepsilon^*,0}(x_0) - R_L^{\mathrm{base},0}(x_0) + \delta \Delta R_L(x_1^0, u_1^*) \\
&= \Delta R_L(x_0, u_0^*) + \delta \Delta R_L(x_1^0, u_1^*).
\end{aligned}$$

Analogously, we can show that for any $k = 1, 2, ...$

$$\begin{aligned}
&J_\infty^{\mu^*,k}(x_0) - J_{\mathrm{base}}(x_0) \\
&= \Delta R_L^{\mu_\varepsilon^*}(x_0) + \delta \Delta R_L^{\mu_\varepsilon^*}(x_1^1) + ... + \delta^k \Delta R_L^{\mu_\varepsilon^*}(x_k^k) \\
&= \Delta R_L(x_0, u_0^*) + \sum_{t=1}^{k} \delta^t \Delta R_L(x_t^{t-1}, u_t^*) \geq 0,
\end{aligned}$$

where x_t^{t-1} is the state after implementing the optimal policy μ_ε^* for t consecutive steps and starting from $x(0) = x_0$. Given the boundedness of the involved functions and a discount factor $\delta < 1$, the conclusion follows by taking the limit as $k \to \infty$. □

Proposition 5.1 guarantees that, when an optimal policy is designed on the basis of the 0-plus-L-step-control utility and it is implemented repeatedly, then the infinite-step utility may not decrease. Such observation simplifies significantly the design process of an optimal policy. In particular, note the following:

- Given the finite length of L steps ahead, based on which the 0-plus-L-step-control is defined, there are no long-term impacts on the utility. Thus, the main challenges of this optimization discussed in Section 3.4 have indirectly been addressed.
- Note that it is sufficient to accurately approximate the 0-plus-L-step-control utility difference $\Delta R_L(x, \mu)$ in order to design an optimal policy. This provides an additional computational advantage compared to standard ADP approaches, where an approximation is usually performed on the generic infinite-step utility J_∞^μ.

5.3 Zero-step-control and arbitrage

The selection of the L-step horizon, over which policies/controls are computed, is based on Assumption 5.1. The size of L corresponds to the time within which the battery cannot reach both its two extreme states (fully charged and empty). For standard batteries the future horizon of L steps will correspond to 4-5 hours, which is a rather short optimization horizon as compared to the one-day optimization horizon. The question that naturally emerges is whether such myopic decision-making process neglects significant profit opportunities. Additional profit opportunities could have been generated through arbitrage, when charging the battery at times when the price is low, and discharging when the price is higher. Such reasoning though would neglect the long-term impacts on the baseline operation of such actions (as we discussed in Section 3.4). Even if we neglect such long-term impacts on the baseline operation, arbitrage could never be profitable for the following reasons: a) any energy procurement is additionally charged with grid tariffs (which are fixed and usually in the range of 60 Euros / MWh); b) any extra battery charge/discharge (outside the baseline needs) incurs additional energy losses; c) in several countries, *green certificates* should be issued for any energy that is fed into the grid. The above limitations render arbitrage practically impossible. As a consequence, we could only gain by appropriately shifting the charging/discharging (baseline) schedules of the battery, and this can always be performed by exploiting the forecast quantities $\phi_i^b(t)$ and $\psi_i^b(t)$, computed over the L-step horizon.

5.4 Zero-step-control approximation

Given the guarantees of Proposition 5.1, it remains to provide an approximation function for the zero-step-control utility difference $\Delta R_L(x, \mu(x))$. We introduce the following approximation $\varrho(x, u)$, which is defined as follows:

$$\begin{aligned}&\varrho(x_t, u_t) \doteq \\ &\alpha_1 \cdot [u_t v_{c,i}(t)]_+ \cdot [\eta_{c,i}\eta_{d,i}\lambda_{\mathrm{DA},f}(t) - \lambda_{\mathrm{DA}}(t)]_+ - \\ &\alpha_2 \cdot [u_t v_{d,i}(t)]_+ \cdot [\lambda_{\mathrm{DA},f}(t) - \eta_{c,i}\eta_{d,i}\lambda_{\mathrm{DA}}(t)]_- + \alpha_3. \quad (8)\end{aligned}$$

The control input u_t is given by (6) where we replace ΔR_L with $\varrho(x_t, u_t)$. The parameters α_1, α_2 and α_3 are unknown and need to be estimated. We impose the additional constraint that $\alpha_1 > 0$ and $\alpha_2 > 0$. The approximation function $\varrho(x, u)$ can be trained using standard linear regression (with a non-linear basis functions). This approximation function tries to generate profitable shifts of procurement and feed-in times. For example, if the baseline operation of a household i in the upcoming short-term horizon of L steps procures energy from the grid, then we may consider procuring this energy earlier if the price satisfies $\eta_{c,i}\eta_{d,i}\lambda_{\mathrm{DA},f}(t) - \lambda_{\mathrm{DA}}(t) > 0$.

The specific choice of ϱ function can be used to generate actions in place of ΔR_L in (6). It generates policies that are specifically tailored to increase the utility within the L-step future horizon. Let $\bar{\mu}$ be the resulting policy based on the approximation ϱ. It is straightforward to check that if the accuracy of the approximation satisfies $\|\varrho(x, \bar{\mu}(x)) - \Delta R_L(x, \mu_\varepsilon^*(x))\| \leq \theta$ uniformly on x, then the performance guarantees of Proposition 5.1 are satisfied as long as $\theta < \varepsilon$.

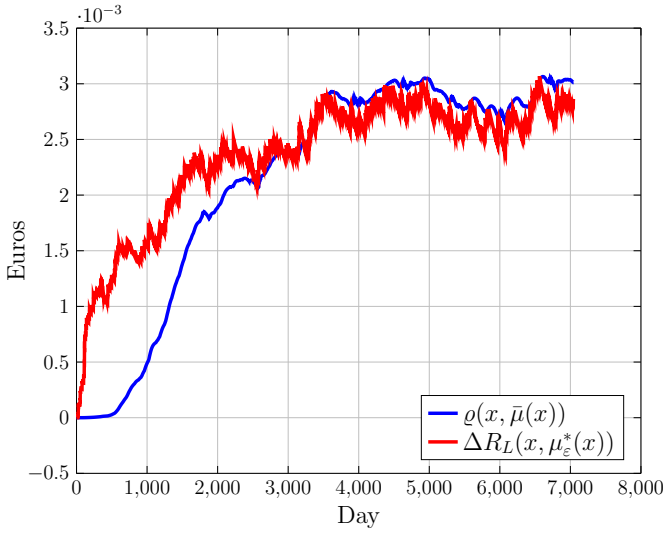

Fig. 4. Approximation of the zero-step-control utility difference.

6. EVALUATION

To evaluate the proposed framework, we performed simulations on real-world data collected from $N = 30$ battery-storage systems located in the state of Upper-Austria, over the duration of approximately one year. According to such simulations, at the beginning of each day, we generate actions over the 96 intervals of the following day (by employing the approximation function $\varrho(x, u)$ and using forecast data). We implement this sequence of actions and at the end of the day we perform a training update of function ϱ given the actual utility recorded. Given that one year of data is usually not enough, we used these data more than once. A small subset was reserved for testing.

Figure 4 provides the performance of the training process, and shows that indeed function ϱ approximates well the zero-step-control utility difference. Figure 5 provides a sample response of the controller when generating the optimal commitment levels for the next day and for a single household. In the first figure, we see the commitment decisions (black line) which are positive when charging and negative when discharging. In the second figure, we see the impact of the commitments on the total and baseline utility. Finally, in the last figure, we see how the state-of-charge evolves. Note that the decision, based on the zero-step-control approximation, was to shift the discharging of the battery at an earlier time, when the price was higher.

Finally, Figure 6 provides a comparison of the proposed zero-step-control ADP methodology with the corresponding solution under a Linear-Programming formulation of the problem and over one week. We observe that the overall credit created by the two methods is almost identical. Furthermore, we also observe that indeed the zero-step-control ADP methodology quite robustly increases the utility as Proposition 5.1 dictates.

7. CONCLUSIONS

This work presented an optimization framework for optimal participation of an aggregator in the DA spot-market through the direct control of a set of residential battery-

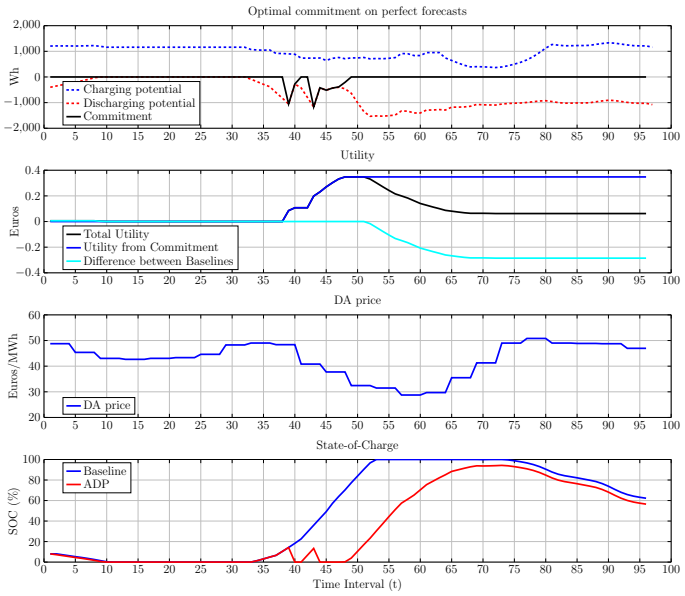

Fig. 5. Sample response of zero-step-control ADP over one day and for a single household.

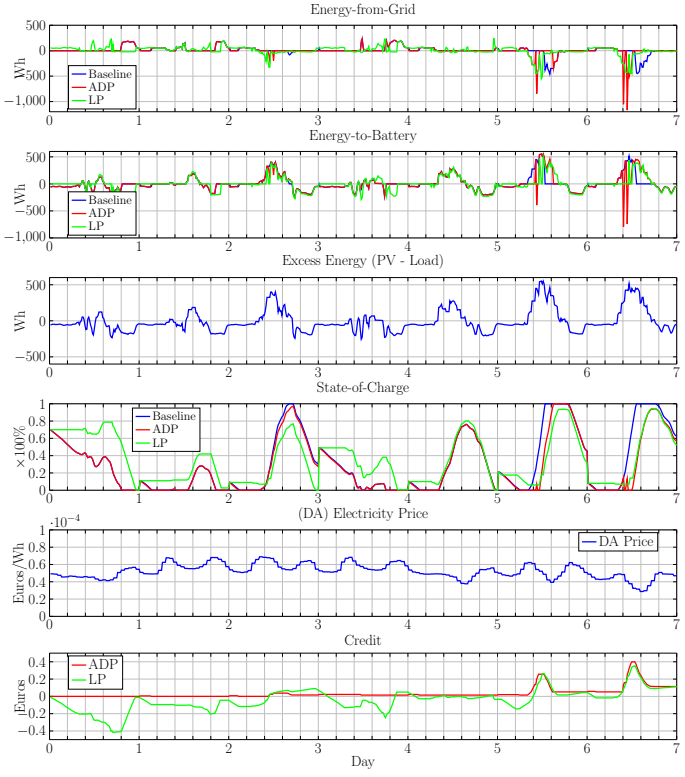

Fig. 6. Comparison between zero-step-control ADP and LP solutions.

storage systems. The aggregator optimizes the amount of flexibility (energy that can be charged/discharged in the participating batteries) that can be offered to the DA market during the next day. Given the possibility of erroneous forecasts, as well as the complexity of the involved optimization, we proposed a reinforcement-learning methodology that trains over time and provides a control strategy for each time interval of the next day. The derivation was based on a zero-step-control approximate-dynamic-programming architecture that takes advantage of the specifics of this problem, and can provide analytical guarantees with respect to the derived long-term performances.

REFERENCES

Chasparis, G.C., Pichler, M., Spreitzhofer, J., and Esterl, T. (2019). A cooperative demand-response framework for day-ahead optimization in battery pools. *Energy Informatics*, 2, 1–17.

Chen, C., Wang, J., and Kishore, S. (2014). A Distributed Direct Load Control Approach for Large-Scale Residential Demand Response. *IEEE Transactions on Power Systems*, 29(5), 2219–2228.

Gomez-Villalva, E. and Ramos, A. (2003). Optimal energy management of an industrial consumer in liberalized markets. *IEEE Trans. Power Syst.*, 18(2), 716–723.

He, G., Chen, Q., Kang, C., Pinson, P., and Xia, Q. (2016). Optimal Bidding Strategy of Battery Storage in Power Markets Considering Performance-Based Regulation and Battery Cycle Life. *IEEE Trans. Smart Grid*, 7(5), 2359–2367.

Herter, K. (2007). Residential implementation of critical-peak pricing of electricity. *Energy Policy*, 35(4), 2121–2130.

Iria, J.P., Soares, F.J., and Matos, M.A. (2017). Trading small prosumers flexibility in the day-ahead energy market. In *2017 IEEE Power & Energy Society General Meeting*, 1–5. Chicago, IL.

Jiang, D.R. and Powell, W.B. (2015). Optimal Hour-Ahead Bidding in the Real-Time Electricity Market with Battery Storage Using Approximate Dynamic Programming. *INFORMS Journal on Computing*, 27(3), 525–543.

Kairies, K., Haberschusz, D., Ouwerkerk, J., Strebel, J., Wessels, O., Magnor, D., Badeda, J., and Sauer, U. (2016). Wissenschaftliches mess-und evaluierungsprogramm solarstromspeicher 2.0. jahresbericht 2016.

Mohsenian-Rad, H. (2016). Optimal Bidding, Scheduling, and Deployment of Battery Systems in California Day-Ahead Energy Market. *IEEE Trans. Power Syst.*, 31(1), 442–453.

Nan, S., Zhou, M., and Li, G. (2018). Optimal residential community demand response scheduling in smart grid. *Applied Energy*, 210, 1280–1289.

Parvania, M., Fotuhi-Firuzabad, M., and Shahidehpour, M. (2013). Optimal Demand Response Aggregation in Wholesale Electricity Markets. *IEEE Transactions on Smart Grid*, 4(4), 1957–1965.

Philpott, A. and Pettersen, E. (2006). Optimizing Demand-Side Bids in Day-Ahead Electricity Markets. *IEEE Trans. Power Syst.*, 21(2), 488–498.

Ruiz, N., Cobelo, I., and Oyarzabal, J. (2009). A Direct Load Control Model for Virtual Power Plant Management. *IEEE Transactions on Power Systems*, 24(2), 959–966.

Triki, C. and Violi, A. (2009). Dynamic pricing of electricity in retail markets. *4OR*, 7(1), 21–36.

Xu, Y., Li, N., and Low, S.H. (2016). Demand Response With Capacity Constrained Supply Function Bidding. *IEEE Transactions on Power Systems*, 31(2), 1377–1394.